Who are you, you who speak? Transducer cascades for information retrieval

Denis Maurel, Nathalie Friburger, Iris Eshkol

Who are you, you who speak?
Transducer cascades for information retrieval

Denis MAUREL[1], Nathalie FRIBURGER[1], Iris ESHKOL[2]

[1]Université François Rabelais Tours, LI
[2]Université d'Orléans, LLL
{denis.maurel, nathalie.friburger}@univ-tours.fr, iris.eshkol@univ-orleans.fr

Abstract

This paper deals with a survey corpus. We present information retrieval about the speaker. We used finite state transducer cascades and we present here detailed results with an evaluation.

Keywords: Information retrieval, Named entity task, Transducer cascades, Survey

1. Motivation

This work is part of a French project to enhance the corpus ESLO (sociolinguistic survey taken in the city of Orléans). This survey has been realized in 1968 and the project is to save records in computer format, to transcribe them and to increase the transcription with annotations in XML format. This work was supported by a French ANR contract (ANR-06-CORP-023) and by European fund from *Région Centre* (FEDER).

The corpus represent a collection of 200 interviews with the questions about the life in the city of Orléans: *How long have you lived in Orléans for?*, *What led you to live in Orléans?*, *Do you like living in Orléans?*, etc. and questions about the occupation or the family of the speaker, completed by recordings within a professional or private context. The recording situations are different: interviews, discussions between friends, recordings in microphone hidden, interviews with the political, academic and religious personalities, conversations between a social worker and parents in Psycho Medical Center of Orleans. In total, we have 300 hours of speech estimated to 4,500,000 words. More precisely, we worked on almost 120 transcribed hours representing 112 Transcriber XML files and 32 577 Kb. We worked on 105 files (31 004 Kb) and we evaluated the results on 7 files (1 573 Kb - 5.1%). The transcription files have no punctuation marks, but the first letter of proper names is capitalized and acronyms are fully capitalized.

We used the CasSys system (Friburger, Maurel, 2004) that computes texts with transducer cascades (Abney, 1996). The cascades we used are hand built: each transducer describes a local grammar for the recognition of some entities. Some times this recognition needs the succession of two or more transducers, in a specific order.

More precisely, we used two cascades; the first one, for named entity recognition, was built some years ago for a newspaper corpus and we adapted it to oral corpus in the project; the second one aimed at discovering information about the speaker in three domains: origin (*is he/she Orléans city native or where he/she comes from?*), family (*is he/she married, with children or not?*) and occupation (*what is his/her occupation? where does he/she work?*). We called this information *designating entities*. This second cascade was specifically built for the project.

CasSys computes transducers with Unitex software (Paumier, 2003) that needs to segment the text by preprocessing. For written text, this segmentation usually uses sentence boundary detection (Friburger and al., 2000). In our corpus there is no punctuation. So we have chosen to use XML Transcriber tags to do the segmentation and also to hide the inside of the tag for the named entity task, sometimes ambiguous with context entities (Dister, 2007).

2. CasEN, the named entity cascade

Our first cascade, CasEN, recognizes entities and adds XML tags to the text. For instance, the input text:
et le général De Gaulle lui-même…
(and General De Gaulle himself…)

Becomes the output text:
et le<**ENT type="pers.hum"**> général De Gaulle<**/ENT**> lui-même…

2.1. Typology

The types used are the seven Ester types[1], with some additional ones.

pers *(person)*
  pers.hum *(human)*, pers.anim *(animal)*
fonc (post)
  fonc.pol (political), *fonc.mil* (military), *fonc.admi* (administrative), *fonc.rel* (religious), *fonc.ari* (aristocratic)
org (organization)
  org.pol (political), *org.edu* (educational), *org.com* (commercial), *org.non-profit* (non commercial), *org.div* (media and recreation), *org.gsp* (administrative)
loc (location)
  loc.geo (geographical), *loc.admi* (administrative), *loc.line*, *loc.fac* (facilities)
  loc.addr (address), *loc.addr.post*, *loc.addr.tel*, *loc.addr.elec*
prod (product)
  prod.vehicule, *prod.award*, *prod.art*, *prod.doc*
time (date and hour)
  time.date, *time.date.abs* (absolute), *time.date.rel* (relative), *time.hour*
amount
  amount.cur (currency),
  amount.phy (physical)

[1] http://www.afcp-parole.org/ester/

amount.phy.age, *amount.phy.dur* (duration), *amount. phy.temp* (temperature), *amount. phy.len*, *amount.phy.area*, *amount.phy.vol* (volume), *amount. phy.wei*, *amount. phy.spd* (speed), *amount.phy.other*

We added subtypes and one new type: eight person subtypes for the name phrase entity and one for a specific category, the names of dynasty; one type for events, with two subtypes, on one hand historical event and on other hand sporting or cultural event:

pers (person)

pers.hum.tit *(Mr, Mrs…)*, pers.hum.gent *(toponymic adjective)*, pers.hum.occ *(occupation)*, pers.hum.sp *(sports)*, pers.hum.art, pers.hum.nat *(nationality)*, pers.hum.rel *(religious)*, pers.hum.pol *(political)*, pers.hum.fonc *(Dr…)*, pers.hum.dynasty

event

event.hist, *event.manif*

2.2. Cascade and examples

The CasEN cascade uses the Delas dictionary (Courtois, Silberztein, 1990) and ten specific dictionaries with 28 341 first names, 31 580 occupations (Gazeau, Maurel 2006), 3 016 acronyms, 114 511 proper names (and derived words) extracted from Prolexbase (Maurel, 2008), 497 animal names, 296 sport names, 110 currency names, 53 car names and 26 newspaper names. Then the cascade is made up of 152 transducers.

For instance, presents transducer recognition of phrases as *I am 22 years old*, *a 22 years old man*, *a woman about twenty years old*, etc. and also the date *from 22 years*.

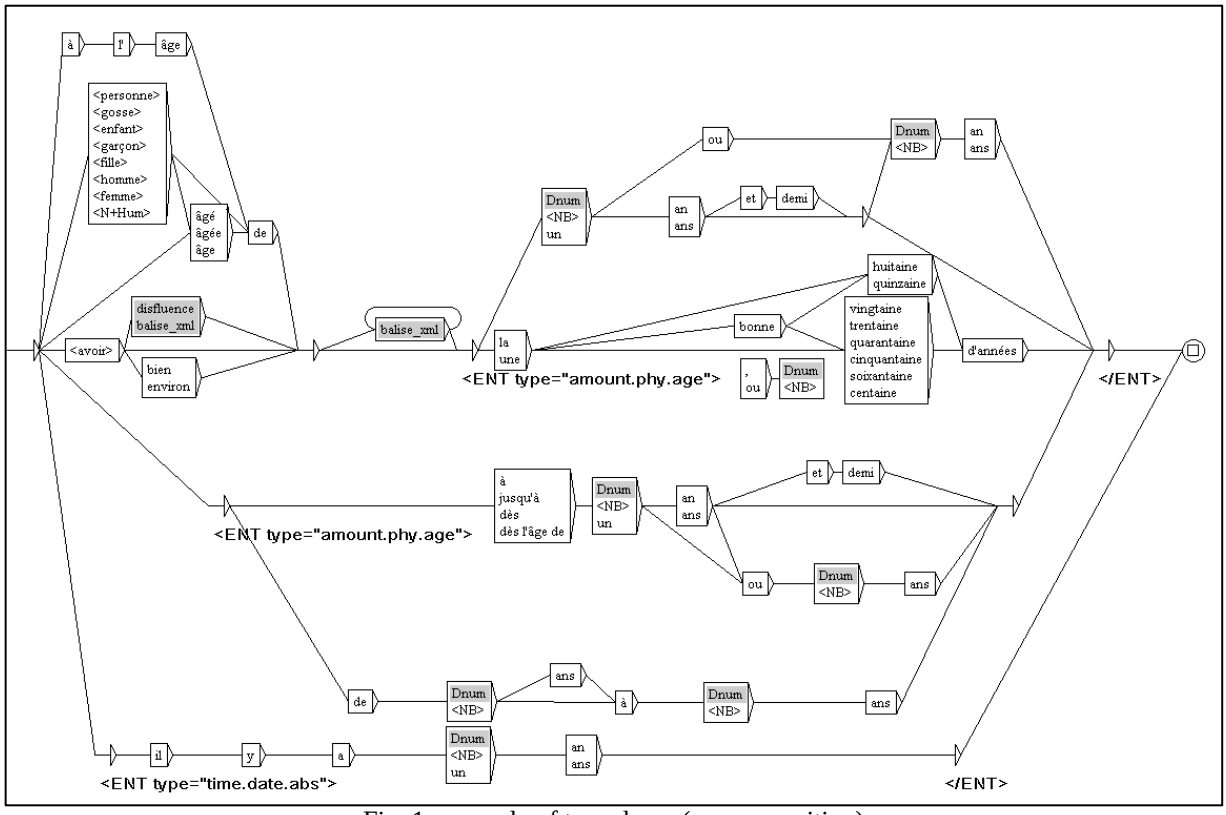

Fig. 1: example of transducer (age recognition)

For instance, we recognized and tagged:

<ENT type="pers.hum">abbé Cartier**</ENT>** se
(Father Cartier)
le **<ENT type="fonc.pol">**président de la république **</ENT>**
(President of the Republic)
le **<ENT type="org.pol">**ministère de l'Education Nationale**</ENT>**
(Ministry of Education)
au **<ENT type="org.edu">**lycée de Montluçon**</ENT>** qui
(secondary school of Montluçon)
des gars de d'**<ENT type="org.com">**EDF**</ENT>** ou qui
(some guys from the EDF Company)
de **<ENT type="org.div">**France Inter**</ENT>** c'est pas ça
(France Inter radio)
la **<ENT type="loc.fac">**piscine d'Orléans**</ENT>** mais
(swimming pool of Orléans)
oui c'est le **<ENT type="loc.geo">**bassin d'Arcachon**</ENT>**
(Arcachon basin)
l'**<ENT type="loc.admi">**Allemagne de l'Est**</ENT>** euh
(East Germany)
qui était **<ENT type="loc.line">**place De Gaulle**</ENT>**
(De Gaulle square)
une **<ENT type="prod.art">**encyclopédie Quillé**</ENT>** j'ai
(Quillé encyclopedia)
le **<ENT type="prod.doc">**journal d'Orléans**</ENT>**
(Orléans newspaper)
<ENT type="time.date">en dix-neuf cent trente-huit**</ENT>**
(in 1938)
oh **<ENT type="time.date">**il y a quinze jours**</ENT>**
(fifteen days ago)

<ENT type="time.hour">jusqu'à sept heures et demie**</ENT>**
(till half past seven)
<ENT type="amount.cur">cinquante mille franc**</ENT>** de
(50 000 French francs)
au moins **<ENT type="amount.phy.wei">**deux cent cinquante kilos**</ENT>** de
(250 kg)

2.3. Evaluation

We present now the evaluated results with precision and recall on the seven test files. Mistakes are mainly due to ambiguities, hesitancies or repetitions of the speaker, transcription mistakes and unexpected XML tags… And omissions where we wrote the rules…

There are 1 305 entities; we recognized 1 227 entities; 27 entities were mistakenly recognized and 51 entities were not recognized. Here, we don't take the correctness of the type and the good position of the brackets into account.

Precision	97.8%
Recall	94.0%

Tab. 1: Partially entities without type

For instance, we forgot:
que Babeth le dise
(diminutive of Elisabeth, not in the first name dictionary)

And celebrities without first name (Monet, Rabelais, Renoir…).

Among these entities, we made 73 mistakes about the type; so we recognized and correctly typed 1 154 entities (always without taking the good position of the brackets into account).

Precision	92.0%
Recall	88.4%

Tab. 2: Partially entities with type

For instance, we did not tag correctly:
environ **<ENT type="amount.cur">**cinq mille livres**</ENT>** euh en rayon
(about 5 000 books)

Because the French word *livre* is ambiguous and means *book*, but also *pound*! And below, we tagged *head of choral* with a political tag, because of a too permissive local grammar:
par un **<ENT type="fonc.pol">**chef de chorale**</ENT>**
(head of choral)

Finally, we completely recognized, with the correct tags, 1 142 entities and we partially recognized 12 entities (11 tags too short and 1 tag too long).

Precision	91.1%
Recall	87.5%

Tab. 3: Complete entities with type

For instance, we did not tag correctly (because there is an unexpected space before the dash):
l'ancienne **<ENT type="org.edu">**faculté de droit de Saint**</ENT>** -**<ENT type="loc.admi">**Yves**</ENT>**
(law faculty of Saint-Yves)

At the beginning of the year, we took part in the ESTER campaign about French spoken language transcription. More exactly at the named entity recognition task (detection and categorization) (Galliano et al., 2009). This corpus has a lot of disfluencies…So, we obtained 79.39 of precision and 65.82 of recall.

3. CasDen, the designating entity cascade

Our second cascade did not work on the corpus, but on the named entity tagged corpus. We tried to locate where information about the speaker and what kind of information can designate him/her.

3.1. Typology

The survey essentially deals with the speaker and his/her family : their origin, age, birth, arrival, work, trade union, etc. First, we defined tags for the person who speaks or whom one speaks:
pers *(person)*
pers.speaker *(the speaker)*, pers.spouse *(his/her spouse)*, pers.child *(his/her children)*, pers.parent *(his/her parents).*

Second, we defined tags for identity, work and trade union:
identity
identity.age, identity.origin *(where he/her comes from)*, identity.birth, identity.arrival *(when he/her came at Orléans)*, identity.children.
work
work.occupation, work.field, work.location, work.business
trade union

We possibly added values for number of child, age, etc.

3.2. Information from question

The first information that we searched is the occupation of the speaker. Generally, the interviewer directly asks about it. So, we obtained simple phrases such as *I am…* or *I was…*:
<DE type="pers.speaker"> je suis **<DE type="work.occupation">** postier**</DE></DE>**
(postal worker)

Sometimes, this information is interrupted by Transcriber tags with pauses:
<DE type="pers.speaker"> je suis **<DE type="work.occupation">** boulanger <Sync time="243.566"/> <Sync time="245.347"/> pâtissier**</DE></DE>**
(baker and confectioner)

And sometimes it is necessary to recognize detailed occupation as *teacher of maladjusted child* or *military social worker* or *technical education teacher*.

The second information is also part of an answer to the interviewer: *For how long has the speaker lived in Orléans*. For instance, here, the speaker hesitates between *27* and *1927*:
<DE type="pers.speaker"> eh bien euh j'habite **<ENT type="loc.admi">**Orléans**</ENT>** depuis <Sync time="10.389"/> <Sync time="11.018"/>**<DE type="identity.arrival">** vingt-sept**</DE>** <Sync time="11.796"/> </Turn> <Turn speaker="spk3" startTime="12.096" endTime="12.814"> <Sync time="12.096"/> hm <Sync time="12.305"/> </Turn> <Turn speaker="spk2" startTime="12.814" endTime="14.586"> <Sync time="12.814"/>**<DE type="identity.arrival">** dix-neuf cent vingt-sept**</DE></DE>**
(I have lived in Orléans since 27… 1927)

3.3. Unsolicited information

But information can be lost inside of conversation. We searched family information such as:

Number of children (we added this information inside the tags):

<DE type="pers.speaker"> j'ai**<DE type="identity.children" value="4">** quatre enfants**</DE></DE>**

(I have four children)

Husband, wife, children or parents occupation:

<DE type="pers.spouse">mon mari est **<DE type="work.occupation">** charcutier**</DE></DE>**

(My husband is a pork butcher)

<DE type="pers.parent">mon père était **<DE type="work.occupation">** mécanicien**</DE>** **<DE type="work. business">**aux chemins de fer**</DE></DE></DE>**

(My father was a rail engineer)

and so:

<DE type="pers.parent">mon père était fils d'un d'un**<DE type="work.occupation">** cultivateur**</DE></DE>**

(My father was the son of a farmer)

And some temporal or local information:

<DE type="pers.speaker">je suis né en**<DE type="identity.origin"><ENT type="loc.admi">** Lorraine**</ENT></DE></DE>**

(I was born in the Lorraine region)

It is also possible that the speaker anticipates some questions that are not asked. For instance, his/her occupation, etc.

3.4. Evaluation

For this work, the goal is to locate information and we evaluate this point. Sometimes designating entities are not precisely marked up but they are located: it is what is important for us.

It is difficult to evaluate this work because the information is very sparse. We used of course the same seven test files. There are only 77 designating entities; we just recognized (almost partially) 69 entities; we made 4 mistakes and we failed to recognise 12 entities:

Precision	94.2%
Recall	84.4%

Tab. 4: Designating entities

For instance, we forgot the following entity, because of a preposition repetition:

euh j'habitais dans dans le**<ENT type="loc.admi">** Berry**</ENT>** à**<ENT type="loc.admi">** Bourges**</ENT>**

(I lived in the Berry region in the city of Bourges)

During this evaluation, we realized that the speaker gives his occupation just by using a verb, for instance *to teach* instead of *teacher*:

euh j'enseignais le français

(I taught French)

We added new transducers to recognize these information.

We also made series of mistakes, when named entity recognition has failed. Here, we forgot *financial agency* with the first cascade and then *I work for the financial agency* with the second one:

je travaille actuellement à l'agence financière du **<ENT type="loc.geo">**bassin Loire-Bretagne**</ENT>**

(I currently work for the financial agency of the Loire-Bretagne area)

4. Conclusion and application

This paper presented extraction of speaker information in a survey corpus. Our tags will allow researchers a quick access to speakers' personal information, especially for sociological studies. In this way, it would be possible to get a database with personal information about each speaker. The study of the database reflects a portrait of an urban society for years 68-69. Our tags will allow also the navigation in the corpus following the sociological criteria such as occupation, age of the speaker or the number of years he lived in Orleans. The user of the corpus could see how a particular speaker defines his profession, his city, etc.

We will continue our work publishing the corpus on the Web (at the end of the ANR project). This project implies an approach based on the best legal and ethical practices. (Baude et al., 2006). The people who have been recorded had not given their permission for use of their words. Thus, the diffusion online of the corpus requires his anonymization. These tags will probably also be used to anonymize the corpus.

References

Abney S. (1996), Partial Parsing via Finite-State Cascades, *Workshop on Robust Parsing*, 8th *European Summer School in Logic, Language and Information*, Prague, Czech Republic, 8-15.

Baude, O., *et alii* (2006), *Corpus oraux : guide des bonnes pratiques 2006.* Paris et Orléans, CNRS-Editions et PUO.

Courtois B., Silberztein M. (1990), Dictionnaires électroniques du français, *Langues française*, 87:11-22.

Dister A. (2007). De la transcription à l'étiquetage morphosyntaxique. Le cas de la banque de données textuelles orales VALIBEL. Thèse de linguistique. Université catholique de Louvain.

Friburger N., Dister A., Maurel D. (2000), Améliorer le découpage des phrases sous Intex, *Revue Informatique et Statistique dans les Sciences Humaines*, 36(1-4):181-200.

Friburger N., Maurel D. (2004), Finite-state transducer cascades to extract named entities in texts, *Theoretical Computer Science*, vol. 313, 94-104.

Galliano S., Gravier G., Chaubard L. (2009), The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts, *Interspeech 2009*.

Gazeau M. A., Maurel D. (2006), Un dictionnaire INTEX de noms de professions : quels féminins possibles ?, *Cahiers de la MSH Ledoux*, 115-127.

Maurel D. (2008), Prolexbase. A multilingual relational lexical database of proper names, *Sixth language resources and evaluation conference* (*LREC 2008*), Marrakech, Maroc, 28-30 mai.

Paumier S. (2003), *De la Reconnaissance de Formes Linguistiques à l'Analyse Syntaxique*, Thèse de Doctorat en Informatique, Université de Marne-la-Vallée.